HYBRID-VGGT: SCALING VISUAL GEOMETRY GROUNDED TRANSFORMERS TO HIGH RESOLUTION VIA MULTI-SCALE DETAIL INJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward foundation models have significantly accelerated uncalibrated 3D reconstruction by inferring geometry in a single pass. However, their reliance on global self-attention introduces a severe quadratic computational bottleneck at high resolutions, forcing a compromise between inference efficiency and the preservation of fine-grained geometric details. To address this accuracy-efficiency trade-off, we introduce Hybrid-VGGT, an asymmetric dual-branch extension of VGGT that allocates global multi-view reasoning and high-resolution detail extraction to complementary streams. A low-resolution branch efficiently anchors global scene context and camera prediction, while a parallel Hybrid Detail Extractor (HDE) captures per-view, fine-grained structural cues directly from high-resolution inputs. By integrating multi-layer local convolutions with scalable linear attention, the HDE extracts structural cues across multiple scales. These multi-scale features are then progressively injected into the global low-resolution stream via a customized hierarchical fusion head, effectively refining the coarse geometric representation with high-frequency boundary cues. Comprehensive experiments demonstrate that Hybrid-VGGT achieves state-of-the-art performance on multiple benchmarks for camera pose recovery, dense depth, and point map estimation, yielding high-fidelity 3D geometry and the best overall mean rank in our comparisons. On 100-frame 1036p clips, Hybrid-VGGT runs at 38.56 FPS, compared with 3.12 FPS for VGGT and 5.19 FPS for Pi3.

1 INTRODUCTION

Extracting accurate 3D geometry from uncalibrated multi-view images is critical for real-world applications such as autonomous navigation Sun et al. (2026), augmented reality, and robotics Murai et al. (2025). In these demanding domains, inference efficiency and end-to-end processing capabilities are paramount. While traditional approaches rely heavily on the intricate complexity of Multi-View Stereo (MVS) Cao et al. (2022); Cheng et al. (2024; 2022); Dai et al. (2017); Yao et al. (2020) pipelines or the high computational costs of iterative optimization, the recent emergence of feed-forward foundation models offers a highly efficient alternative. The Visual Geometry Grounded Transformer (VGGT) Wang et al. (2025a) stands as a landmark model in this shift. By directly inferring camera parameters and dense 3D attributes in a single forward pass, VGGT completely bypasses the need for explicit geometric priors and multi-stage processing. It natively unifies the prediction of depth maps, point clouds, and multi-view consistency, delivering highly competitive results across diverse benchmarks with real-time inference capabilities.

While these feed-forward models excel at capturing overall scene layouts, their standard operating resolutions strictly limit their precision at the micro-level. Reconstructing sharp depth boundaries, recovering thin structures, and generating dense point clouds in distant regions fundamentally require the rich, dense spatial sampling that only high-resolution inputs can provide. Yet, scaling standard transformers to these higher resolutions exposes a severe architectural bottleneck. Because these architectures rely heavily on dense global self-attention, their computational cost and memory consumption grow quadratically as the token sequence lengthens. To mitigate this quadratic bottleneck, approaches like LiteVGGT Shu et al. (2025) and FastVGGT Shen et al. (2025) utilize token merging and caching. However, such aggressive compression inevitably discards micro-level

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

information, resulting in a significant loss of geometric precision. Concurrently, methods such as CARVE Xu et al. (2025) and DAGE Ngo et al. (2026) explore dual-stream architectures to divide the representation space, typically dedicating a standard transformer-based secondary branch to process high-resolution inputs. Yet, these transformer branches inherently lack the strong local inductive biases Xiao et al. (2021) necessary to capture sharp structural gradients. Furthermore, they typically process high-resolution features at a single, fixed scale. While complex geometric boundaries naturally vary in size and inherently demand varying receptive fields, the absence of a multi-scale extraction mechanism restricts their capacity to fully recover fine-grained details. This ultimately leads to blurred geometric edges and compromised structural fidelity.

In this paper, we propose a novel asymmetric dual-branch architecture, named Hybrid-VGGT, to resolve the conflict between computational scalability and geometric fidelity. Instead of directly passing high-resolution inputs into a computationally expensive global network, our framework handles global scene context and structural details through two dedicated streams. A standard-resolution main branch efficiently establishes robust multi-view coherence. Simultaneously, a concurrent high-resolution auxiliary branch employs our proposed Hybrid Detail Extractor (HDE) to explicitly recover fine-grained boundaries. Specifically, the HDE combines the localized precision of multi-layer convolutions with the expanded receptive field of linear attention, yielding rich high-frequency features across varying scales. Finally, these parallel representations are synthesized via a customized Dense Prediction Transformer head. This module acts as a hierarchical bridge, utilizing the multi-scale boundary information from the HDE to precisely guide the upsampling of global features, ultimately generating sharp 3D geometry without incurring a heavy computational burden.

Extensive experiments validate our architecture and fusion mechanisms. Hybrid-VGGT achieves state-of-the-art performance across rigorous benchmarks for camera pose recovery, dense depth estimation, and 3D point map reconstruction, yielding high-fidelity 3D geometry with highly competitive inference throughput. Our main contributions are summarized as follows:

- We propose Hybrid-VGGT, an asymmetric dual-branch architecture that overcomes the quadratic computational bottleneck of high-resolution 3D reconstruction. By processing global scene context and high-frequency details in parallel streams, it achieves highly efficient inference without compromising geometric fidelity.
- We introduce the Hybrid Detail Extractor (HDE) and a customized Dense Prediction Transformer (DPT) head. The HDE integrates multi-layer convolutions with linear attention to extract precise structural boundaries across multiple scales. These high-frequency cues are then progressively injected into the global stream to recover sharp geometric edges.
- We achieve state-of-the-art performance across rigorous benchmarks for camera pose recovery, dense depth estimation, and 3D point map reconstruction. Qualitative visualizations demonstrate our effectiveness in recovering thin structures and sharp geometric boundaries, yielding high-fidelity 3D geometry alongside highly competitive inference throughput.

2 Related Works

2.1 Feedforward 3D Foundation Models

Extracting geometric information has historically relied on complex multi-stage pipelines like Multi-View Stereo (MVS) Cao et al. (2022); Cheng et al. (2024); Seitz et al. (2006) and Bundle Adjustment Triggs et al. (1999). Recent end-to-end learning methods Long et al. (2024); Lu et al. (2025); Reizenstein et al. (2021); Smart et al. (2024); Zhang et al. (2024) have fundamentally revolutionized this paradigm. Notably, DUSt3R Wang et al. (2024) pioneered the direct regression of view-consistent 3D point maps without camera calibration. This uncalibrated framework was subsequently refined by MASt3R Murai et al. (2025) with robust local feature matching, and significantly accelerated by Fast3R Yang et al. (2025) for rapid multi-view inference.

VGGT Wang et al. (2025a) jointly predicts camera parameters, depth maps, and point clouds in a single forward pass. Pi3 Wang et al. (2025d) introduces permutation-equivariant mechanisms, while the Depth Anything series Yang et al. (2024); Lin et al. (2025) learns robust geometric priors from large-scale datasets.

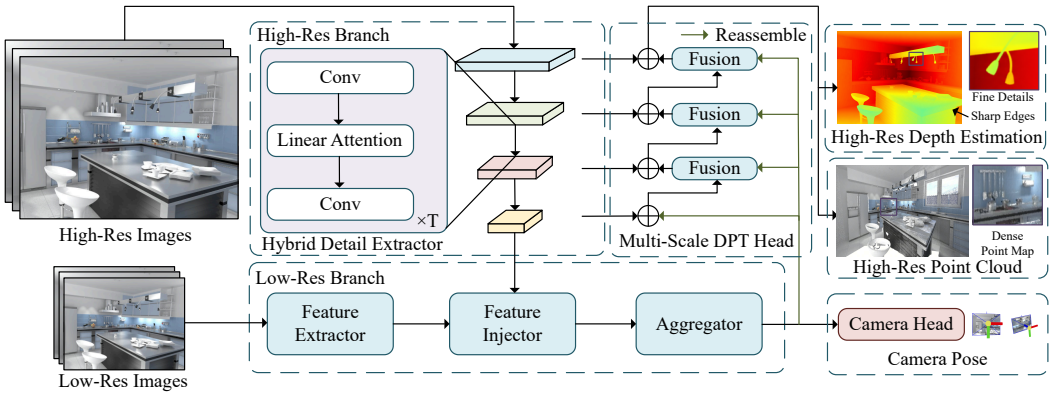


Figure 1: Architectural pipeline of our proposed Hybrid-VGGT. Our framework explicitly separates global geometry modeling and local detail extraction into two dedicated streams. A low-resolution main branch establishes robust global context and camera poses, while a concurrent high-resolution branch employs the Hybrid Detail Extractor (HDE) to capture fine-grained multi-scale structural cues without the quadratic cost of self-attention. These parallel representations are subsequently integrated via a customized Dense Prediction Transformer (DPT) head, which progressively injects high-frequency boundaries into the global stream.

Despite their powerful geometric reasoning, these models are bottlenecked by the quadratic complexity of self-attention at high resolutions. While methods Feng et al. (2025); Yuan et al. (2026) like LiteVGGT Shu et al. (2025) and FastVGGT Shen et al. (2025) mitigate this via token compression, they inevitably discard fine-grained structural cues, leading to blurred edges. Dual-stream architectures such as CARVE Xu et al. (2025) and DAGE Ngo et al. (2026) also process high-resolution inputs separately. Within this design family, our contribution is the integration of convolutional and linear-attention detail extraction with multi-scale injection into VGGT geometry features.

2.2 High-Resolution Depth Estimation

Preserving sharp boundaries and intricate structures in 3D reconstruction inherently demands high-resolution spatial sampling. Recent foundation models like Depth Anything V2 Yang et al. (2024) and MoGe2 Wang et al. (2025c) establish robust geometric priors via massive scale synthetic data and hierarchical optimization. However, scaling their standard transformer backbones to high resolutions remains computationally prohibitive due to the quadratic complexity of self-attention.

To circumvent this, existing high-resolution strategies Fu et al. (2024); Garcia et al. (2025); Ke et al. (2024); Gui et al. (2025); Pham et al. (2025) often impose severe compromises. Patch-wise networks like DepthPro Bochkovskii et al. (2024) improve local fidelity but restrict the global receptive field, introducing stitching artifacts. Conversely, generative approaches using cascaded diffusion priors Li et al. (2022) hallucinate pixel-perfect details but incur massive overhead unsuitable for rapid feed-forward inference. Furthermore, implicit methods like InfiniDepth Yu et al. (2026) offer arbitrary-resolution querying but struggle to reintegrate localized inductive biases into a globally consistent stream. Recognizing these limitations, Hybrid-VGGT addresses these conflicting demands through an asymmetric dual-branch framework. By integrating multi-layer convolutions with scalable linear attention, we extract fine-grained structural details from full-resolution inputs and progressively inject them into the global context, ultimately achieving sharp 3D geometry without sacrificing robust multi-view coherence or real-time inference efficiency.

3 Method

Hybrid-VGGT uses an asymmetric dual-branch architecture to recover high-fidelity geometry from high-resolution inputs while preserving sharp structural boundaries and avoiding dense global self-attention on the high-resolution grid. The design allocates global multi-view reasoning and local detail extraction to complementary streams, allowing dense predictions to incorporate high-

resolution cues without increasing the spatial resolution of the global attention grid. Fig. 1 provides an overview.

3.1 Overall Architecture

Given a sequence of S uncalibrated high-resolution images denoted as $\mathbf{I}_{high} \in \mathbb{R}^{S\times 3\times H\times W}$, we strategically allocate the geometry estimation process across two complementary streams.

First, a Low-Resolution Branch processes a downsampled sequence, $\mathbf{I}_{low} \in \mathbb{R}^{S\times 3\times \frac{H}{2}\times \frac{W}{2}}$. This branch acts as the global geometric anchor, robustly extracting multi-view coherence, scene context, and relative camera poses with minimal computational overhead. Concurrently, a High-Resolution Branch processes the uncompressed sequence $\mathbf{I}_{high}$ using our proposed Hybrid Detail Extractor (HDE). This branch is exclusively dedicated to retrieving fine-grained geometric details and sharp structural boundaries, strictly bypassing costly global attention mechanisms.

Finally, the parallel feature streams from these two branches are progressively integrated through a customized Hierarchical Feature Fusion module (Multi-Scale DPT Head). This injection mechanism elegantly guides the upsampled global geometry features using localized, multi-scale boundary cues from the HDE. The framework ultimately yields high-fidelity dense predictions, including high-resolution depth maps $\mathbf{D}$ and dense point clouds, alongside accurate 6-DoF camera poses.

3.2 Low-Resolution Global Geometry Stream

Establishing multi-view consistency requires a broad receptive field but not micro-level precision. Thus, we process the downsampled sequence $\mathbf{I}_{low}$ through a ViT encoder Dosovitskiy et al. (2020). The extracted tokens are refined by a Feature Injector and a multi-view Aggregator, which employs coupled intra- and inter-view attention to rigorously model epipolar constraints, yielding unified global features $\mathbf{F}_g$. Operating exclusively on $\mathbf{I}_{low}$ circumvents the severe computational bottlenecks of high-resolution processing. Ultimately, $\mathbf{F}_g$ serves a dual purpose: it anchors the structural foundation for subsequent dense reconstruction and routes directly to an auxiliary Camera Head to regress intrinsic matrices and 6-DoF extrinsic poses.

3.3 Scalable High-Resolution Detail Extraction

Complementing the low-resolution global stream, the Hybrid Detail Extractor (HDE) captures fine-grained structural cues directly from uncompressed high-resolution inputs. It targets thin structures and sharp geometric boundaries while avoiding the quadratic attention cost of standard vision transformers. Local convolutions capture spatial neighborhoods, while linear attention supplies a broader receptive field within each view. Intermediate features provide complementary detail cues at multiple scales for the subsequent hierarchical fusion with global geometry features.

3.3.1 The Resolution-Complexity Bottleneck

In standard visual geometry transformers, self-attention Vaswani et al. (2017) operates on a sequence of tokens $\mathbf{X} \in \mathbb{R}^{N\times d}$, where N is the sequence length and d is the embedding dimension. The canonical dot-product attention is formulated as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times d_k}$ denote the query, key, and value matrices. Computing this attention matrix incurs $\mathcal{O}(N^2)$ computational costs. At high resolutions, the sequence length N grows quadratically, rendering this $\mathcal{O}(N^2)$ bottleneck intractable for end-to-end geometric inference.

3.3.2 Formulating the Hybrid Detail Extractor

To circumvent this pervasive bottleneck, we propose a scalable alternative that captures high-frequency structural priors. Our HDE processes the uncompressed high-resolution images $\mathbf{I}_{high}$ through a cascade of hybrid blocks. Let $\mathbf{F}_{in}^{(l)}$ denote the input feature map at the l-th HDE layer. The sequential processing within a single block is mathematically defined as follows.

Local Inductive Bias Injection. To explicitly capture high-frequency boundaries and fine-grained textures, we first inject essential local inductive biases into the feature space. Let $\mathrm{LN}(\cdot)$ denote Layer Normalization and $\mathrm{DWConv}(\cdot)$ represent the depth-wise convolution operation. The localized feature aggregation is formulated as:

$$\mathbf{F}_{conv}^{(l)} = \mathrm{DWConv}\left(\mathrm{LN}(\mathbf{F}_{in}^{(l)})\right) + \mathbf{F}_{in}^{(l)} \tag{2}$$

Scalable Linear Attention. To efficiently approximate global receptive fields, we replace the canonical softmax attention with a linear attention variant Shen et al. (2021). By defining a feature map $\phi(\cdot)$ that acts as a positive activation function, the linear attention mechanism is expressed as:

$$\mathbf{F}_{attn}^{(l)} = \frac{\phi(\mathbf{Q})\left(\phi(\mathbf{K})^\top \mathbf{V}\right)}{\phi(\mathbf{Q})\sum_{j=1}^{N}\phi(\mathbf{K}_j)^\top} \tag{3}$$

By utilizing the associative property of matrix multiplication, we compute the right-side term $\left(\phi(\mathbf{K})^\top \mathbf{V}\right)$ first. This crucial reordering reduces the computational complexity from $\mathcal{O}(N^2 d)$ to $\mathcal{O}(Nd^2)$. Given that $d \ll N$ for high-resolution inputs, this structural modification ensures linear scalability $\mathcal{O}(N)$ while maintaining broad contextual reasoning.

Convolutional Refinement and Multi-Scale Output. After establishing a broad receptive field via linear attention, the feature map undergoes a concluding depth-wise convolution to ensure smooth spatial transitions and consolidate the high-frequency boundaries. Rather than outputting a single-scale representation, our architecture explicitly constructs a feature pyramid to facilitate hierarchical decoding. By cascading T such HDE blocks and extracting intermediate features at specifically designated layers, the High-Resolution Branch generates a multi-scale stack of fine-grained detail maps, denoted as $\mathbf{F}_h = \{\mathbf{F}_{HDE}^{(1)}, \mathbf{F}_{HDE}^{(2)}, \ldots, \mathbf{F}_{HDE}^{(L)}\}$. These hierarchical representations encapsulate pristine high-frequency geometry priors across various native resolutions, functioning as the essential structural cornerstone for the subsequent Multi-Scale DPT Head.

3.4 Hierarchical Cross-Scale Feature Fusion

While the parallel branches efficiently extract global scene context and high-frequency details respectively, their distinct feature representations must be effectively unified. To achieve this, we introduce a customized Dense Prediction Transformer (DPT) head designed to progressively inject the multi-scale boundary cues $\mathbf{F}_h$ directly into the global geometric stream.

Top-Down Global Context Aggregation. To reconstruct the dense 3D geometry, the global multi-view tokens $\mathbf{F}_g$ from the low-resolution main branch are first mapped into a hierarchical feature pyramid. We employ a top-down pathway equipped with RefineNet blocks Lin et al. (2017) to progressively upsample and integrate these global representations. Let $\mathbf{F}_{global}^{(l)}$ denote the intermediate global feature map at the l-th decoder stage. The baseline top-down aggregation is formulated as:

$$\tilde{\mathbf{F}}_{out}^{(l)} = \phi_{refine}^{(l)}\left(\mathbf{F}_{out}^{(l+1)}, \mathbf{F}_{global}^{(l)}\right) \tag{4}$$

where $\mathbf{F}_{out}^{(l+1)}$ represents the fused output from the preceding lower-resolution stage, and $\phi_{refine}^{(l)}$ denotes the fusion operation of the l-th RefineNet block.

High-Frequency Boundary Injection. To explicitly recover sharp geometric edges without incurring heavy computational overhead, we introduce a direct injection mechanism for the high-frequency structural cues. At each hierarchical stage l, the corresponding multi-scale detail feature $\mathbf{F}_{HDE}^{(l)} \in \mathbf{F}_h$ extracted by the HDE branch is resolution-matched via bilinear interpolation, denoted as $\mathcal{B}(\cdot)$. These aligned high-frequency features are then directly injected into the global geometric stream:

$$\mathbf{F}_{out}^{(l)} = \tilde{\mathbf{F}}_{out}^{(l)} + \mathcal{B}\left(\mathbf{F}_{HDE}^{(l)}\right) \tag{5}$$

This mechanism supplies local detail cues to the geometric representation; whether they correspond to true depth discontinuities is learned through geometric supervision.

By recursively applying this cross-scale injection across all pyramid levels, the final feature representation effectively unifies robust multi-view coherence with sharp structural boundaries. This fused representation is then processed by the final prediction head to generate high-fidelity 3D geometry, achieving accurate dense reconstruction at high resolutions.

3.5 Training Details

Training Objectives. We optimize our framework end-to-end using a joint geometric loss:

$$\mathcal{L}_{total} = \mathcal{L}_{pose} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{pts}\mathcal{L}_{pts} \tag{6}$$

Here, $\mathcal{L}_{pose}$ applies hierarchical weighting across iterative stages to enforce strict epipolar alignment. The dense losses ($\mathcal{L}_{depth}$, $\mathcal{L}_{pts}$) are confidence-aware, decoupling absolute geometric error minimization from confidence optimization. Crucially, both incorporate first-order spatial gradient matching to penalize blurriness, directly preserving the pristine high-frequency boundaries extracted by our high-resolution branch.

Implementation Details. The low-resolution stream is initialized with pre-trained VGGT Wang et al. (2025a) weights, while the HDE branch and Multi-Scale DPT Head are initialized randomly. The input frames for the low-res stream are resized to a maximum dimension of 518 pixels, strictly maintaining the original aspect ratio. Concurrently, the HDE branch processes the high-resolution inputs at a maximum dimension of 1036 pixels. This hierarchical architecture enables the direct output of high-fidelity dense depth and point maps at the full 1036p resolution without post-processing upsampling. We train the entire framework on 64 NVIDIA A800 GPUs by optimizing the joint geometric loss. Comprehensive implementation details are provided in the supplementary material.

Training Data. To ensure robust zero-shot generalization across diverse spatial domains, we train our model on a large-scale aggregation of 16 publicly available datasets. This combined corpus provides extensive coverage of both indoor and outdoor environments, encompassing a wide variety of scenes from noise-free synthetic renderings to complex real-world captures. To mitigate label noise and handle data heterogeneity, we apply rigorous dataset-specific geometric filtering and a probability-weighted sampling strategy. A comprehensive list of the dataset mixture and detailed preprocessing protocols are provided in the supplementary material.

4 Experiment

This section evaluates our method against state-of-the-art approaches across three dense reconstruction tasks to validate its effectiveness and high-fidelity output.

4.1 Camera Pose Estimation

We evaluate predicted camera poses using two distinct evaluation protocols based on the characteristics of the datasets. For the object-centric Co3Dv2 (Test Set) Reizenstein et al. (2021) and scene-level RealEstate10K Zhou et al. (2018) datasets, we report the angular accuracy (AUC30 ↑). For the TUM-dynamics Zhao et al. (2022) and ScanNet++ Yeshwanth et al. (2023) benchmarks, we measure distance errors, including the Absolute Trajectory Error (ATE ↓) alongside Relative Pose Errors for both translation (RPE-t ↓) and rotation (RPE-r ↓).

Table 1: Inference speed FPS on 100-frame clips using an RTX 5090.

Resolution	VGGT	Pi3	Ours
518p	44.69	56.11	40.09
1036p	3.12	5.19	**38.56**

As reported in Table 2, our method establishes SOTA results on both Co3Dv2 and TUM-dynamics. Furthermore, it delivers highly competitive performance on par with Pi3 on RealEstate10K and the in-domain ScanNet++ dataset. Crucially, our framework secures the best overall mean rank of 1.62 across all evaluated metrics. This consistent performance across diverse benchmarks indicates the model's robust generalization capabilities, validating the practical advantage of retaining fine-grained geometric priors through our native high-resolution representations.

4.2 Multi-view Depth Estimation

Following CUT3R Wang et al. (2025b), we evaluate multi-view depth (Abs Rel ↓, $\delta < 1.25$ ↑) on Sintel Bozic et al. (2021), ETH3D Schops et al. (2017), 7-Scenes Shotton et al. (2013), BlendedMVS Yao et al. (2020), and ScanNet++ Yeshwanth et al. (2023) using per-sequence scale alignment. Depth metrics are computed on a common 1036p evaluation grid: DAGE and our method output at 1036p, while the other methods use upsampled lower-resolution predictions. This compares the

Table 2: Quantitative comparison of Camera Pose Estimation. The best and second-best results are highlighted with red and orange backgrounds, respectively. The Rank represents the unweighted average ranking across all evaluated metrics (lower is better).

Method	Co3Dv2	RE10K	TUM			Scannet++ (Seen)			Rank ↓
	AUC30 ↑	AUC30 ↑	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓	
Fast3R Yang et al. (2025)	73.43	61.68	0.090	0.101	1.425	0.155	0.123	3.491	6.00
CUT3R Wang et al. (2025b)	75.82	81.47	0.047	0.015	0.451	0.094	0.022	0.629	4.25
FLARE Zhang et al. (2025)	73.99	80.01	0.026	0.013	0.475	0.064	0.023	0.971	4.50
VGGT Wang et al. (2025a)	88.59	77.62	0.012	0.010	0.311	0.035	0.015	0.382	2.63
Pi3 Wang et al. (2025d)	88.41	85.90	0.014	0.009	0.312	0.031	0.013	0.347	1.75
Ours	90.26	84.94	0.011	0.009	0.308	0.033	0.014	0.372	1.62

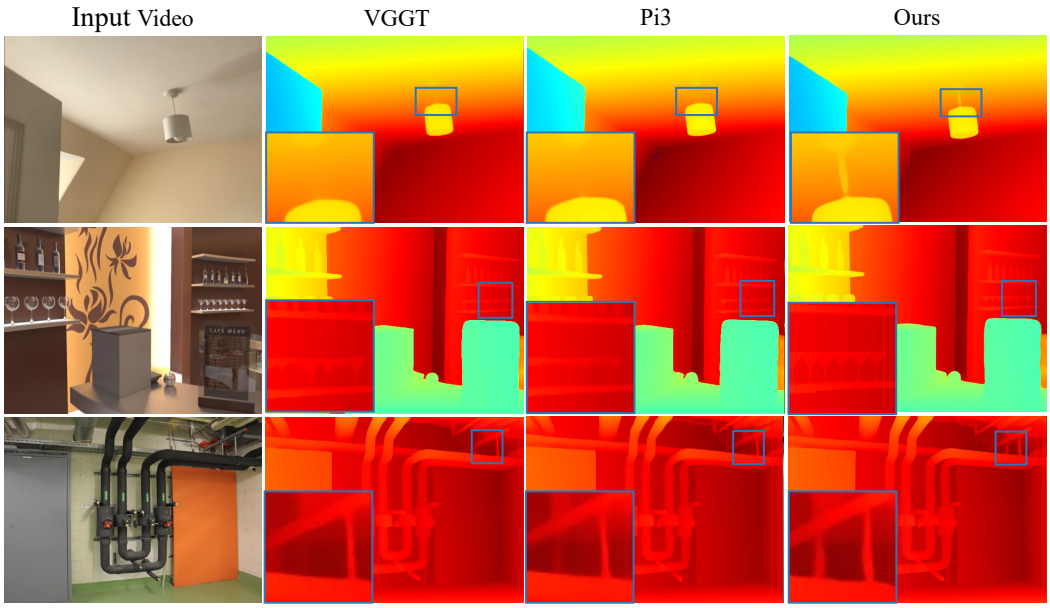


Figure 2: Qualitative comparison of multi-view depth estimation. Our method natively recovers sharp boundaries and intricate thin structures compared with VGGT and Pi3.

complete systems and does not isolate architecture gains from input-resolution gains. As shown in Table 3, our method establishes new state-of-the-art results. Compared against both conventionally upsampled baselines and recent native high-resolution architectures Ngo et al. (2026), our framework demonstrates superior geometric fidelity by securing first place in 7 out of 10 metrics.

Qualitative comparisons shown in Fig. 2 further corroborate this advantage. Unlike the upsampled VGGT and Pi3 baselines, our natively decoded outputs successfully preserve sharp boundaries and intricate thin structures, including lamp cords and distant shelves, effectively eliminating severe over-smoothing artifacts.

Beyond geometric accuracy, our dual-branch architecture maintains high computational efficiency at high resolutions. We benchmark the inference speed in Frames Per Second (FPS) on 100-frame clips using a single RTX 5090 GPU, as detailed in Table 1. Due to quadratic attention bottlenecks, the throughput of VGGT and Pi3 drops significantly at 1036p, yielding 3.12 and 5.19 FPS, respectively. In contrast, our framework operates at 38.56 FPS. This demonstrates that our method can achieve high-fidelity dense 3D reconstruction while maintaining real-time inference capabilities.

4.3 Point Map Estimation

We evaluate unprojected 3D point maps on the object-centric DTU Jensen et al. (2014) and scene-level 7-Scenes Shotton et al. (2013) and ETH3D Schops et al. (2017) datasets. Following standard

Table 3: Quantitative comparison of Multi-view Depth Estimation. The best and second-best results are highlighted with red and orange backgrounds, respectively. Abs Rel is lower-is-better (↓), while $\delta < 1.25$ is higher-is-better (↑). Notably, DAGE and our method natively output at 1036p, whereas other baselines are bilinearly upsampled to 1036p for evaluation.

Method	Sintel		ETH3D		7-Scenes		BlendedMVS		Scannet++ (Seen)		Rank ↓
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	
Dust3R	0.456	0.509	0.241	0.620	0.412	0.914	0.147	0.863	0.145	0.813	6.4
Mast3R	0.380	0.533	0.152	0.804	0.442	0.921	0.166	0.878	0.043	0.974	5.2
Fast3R	0.547	0.429	0.251	0.556	0.506	0.733	0.171	0.826	0.183	0.718	8.6
CUT3R	0.630	0.459	0.154	0.760	0.414	0.959	0.150	0.886	0.031	0.965	5.4
FLARE	0.496	0.418	0.232	0.630	0.480	0.852	0.212	0.820	0.146	0.800	8.0
VGGT	0.226	0.680	0.036	0.986	0.379	0.946	0.020	0.992	0.061	0.962	3.2
Pi3	0.231	0.674	0.045	0.986	0.417	0.966	0.047	0.973	0.047	0.974	3.6
DAGE	0.214	0.695	0.034	0.979	0.423	0.971	0.030	0.988	0.026	0.973	2.8
Ours	0.208	0.685	0.033	0.989	0.368	0.940	0.021	0.993	0.025	0.975	1.6

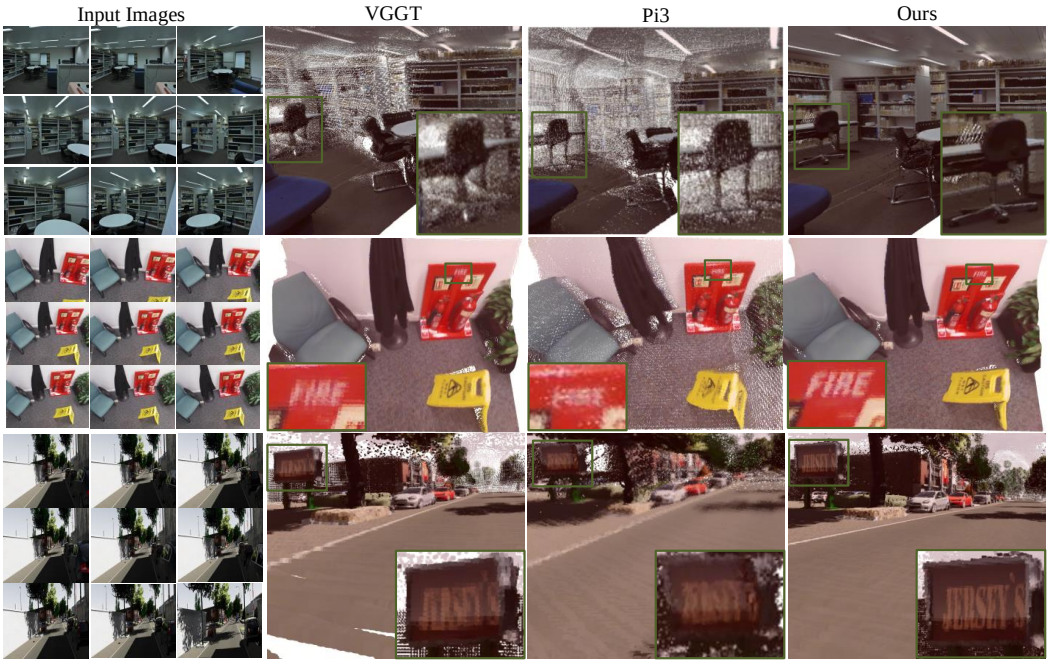


Figure 3: Qualitative comparison of unprojected 3D point maps. We compare our natively decoded high resolution outputs against the VGGT baseline across diverse indoor and outdoor scenes.

protocols, predictions are aligned to ground truth via Umeyama Sim(3) and ICP refinement. We report Accuracy (ACC ↓), Completeness (Comp ↓), Normal Consistency (NC ↑), and the mean Rank across all metrics.

As shown in Table 4, our method achieves a mean Rank of 1.67, significantly outperforming foundation models like VGGT Wang et al. (2025a) and Pi3 Wang et al. (2025d). It leads on all three DTU metrics and on 7-Scenes ACC and Comp, while Pi3 performs better on 7-Scenes NC and all three ETH3D metrics. These results support competitive aggregate performance, with gains that vary across datasets and metrics.

Qualitative results shown in Fig. 3 further corroborate this advantage. Compared to the over-smoothed predictions of VGGT caused by its low-resolution bottleneck, our natively decoded high-resolution outputs successfully preserve legible semantic structures and sharp boundaries. In complex environments, our method yields notably denser and cleaner point clouds with well-preserved planar structures.

Table 4: Quantitative comparison of Point Map Estimation. The best and second-best results are highlighted with red and orange backgrounds, respectively. ACC and Comp metrics are lower-is-better (↓), while NC is higher-is-better (↑). The Rank represents the unweighted average ranking across all evaluated metrics (lower is better).

Method	DTU			7-Scenes			ETH3D			Rank ↓
	ACC ↓	Comp ↓	NC ↑	ACC ↓	Comp ↓	NC ↑	ACC ↓	Comp ↓	NC ↑	
Fast3R Yang et al. (2025)	3.340	2.929	0.671	0.095	0.144	0.673	0.832	0.978	0.667	5.67
CUT3R Wang et al. (2025b)	4.742	3.400	0.679	0.093	0.102	0.704	0.617	0.747	0.754	4.78
FLARE Zhang et al. (2025)	2.541	3.174	0.684	0.085	0.145	0.696	0.464	0.664	0.744	4.11
VGGT Wang et al. (2025a)	1.338	1.896	0.676	0.044	0.056	0.733	0.280	0.305	0.853	2.78
Pi3 Wang et al. (2025d)	1.198	1.849	0.678	0.047	0.075	0.742	0.194	0.210	0.883	2.00
Ours	1.117	1.562	0.696	0.041	0.054	0.729	0.271	0.273	0.850	1.67

4.4 Ablation Study

To explicitly validate our architectural contributions, we conduct an ablation study on the Hypersim Test set Roberts et al. (2021). As shown in Table 5, we evaluate variants across three geometric dimensions: Camera Pose (AUC30 ↑), Depth Error (Abs Rel Mean ↓), and Point Map Error (XYZ Abs Rel Mean ↓).

Table 5: Ablation on core components.

Variant	Pose ↑	Depth ↓	XYZ ↓
Single (L5)	0.9588	0.0230	0.2539
Single (L1)	0.9664	0.0216	0.2484
w/o Injector	0.9642	**0.0215**	0.2387
DINO Ext.	0.9590	0.0226	0.2490
Ours	**0.9682**	**0.0215**	**0.2262**

Effectiveness of Multi-Scale Fusion. Fusing only the highest-resolution features (Single L1) or solely the lowest-resolution features (Single L5) yields suboptimal performance across all metrics. This confirms that aggregating all hierarchical levels is essential to balance broad geometric contexts with fine-grained local boundaries.

The Crucial Role of the Feature Injector. Restricting high-resolution fusion strictly to the final DPT Head (w/o Injector) maintains 2D depth performance but noticeably degrades unprojected 3D point map accuracy, increasing the XYZ error from 0.2262 to 0.2387. This demonstrates that early cross-branch interaction is critical for preserving consistent 3D spatial structures.

Necessity of Dedicated Geometric Extractors. Replacing our Hybrid Detail Extractor (HDE) with a foundational DINO backbone (DINO Ext.) degrades overall performance. Compared with this variant, HDE reduces XYZ error by 9.2%, alongside improvements in both pose and depth estimation. While DINO excels at general semantic representations, our dedicated HDE is specifically optimized to capture the pristine, high-frequency spatial boundaries required for accurate dense 3D reconstruction.

5 Conclusion

In this paper, we present Hybrid-VGGT, a highly efficient architecture designed to scale feed-forward visual geometry models to high resolutions. The core insight of our approach lies in addressing the conflicting computational demands between global multi-view consistency and localized structural sharpness. Specifically, our framework explicitly separates global context aggregation into a standard-resolution stream, while offloading high-frequency detail extraction to our dedicated Hybrid Detail Extractor. This parallel processing strategy successfully circumvents the quadratic computational bottleneck inherent in standard vision transformers. Our hierarchical injection mechanism ensures that these precise, multi-scale boundary cues are seamlessly integrated with the global geometric representation. Extensive evaluations across rigorous benchmarks confirm that Hybrid-VGGT establishes state-of-the-art performance in camera pose recovery, high-fidelity depth, and dense point map estimation while maintaining highly competitive inference efficiency.

REFERENCES

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pp. 690–708. Springer, 2022.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision*, pp. 247–263. Springer, 2024.

Aleksei Bochkovskii, AmaãĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth. *arXiv preprint arXiv:2208.02541*, 2022.

Junda Cheng, Wei Yin, Kaixuan Wang, Xiaozhi Chen, Shijie Wang, and Xin Yang. Adaptive fusion of single-view and multi-view depth for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10138–10147, 2024.

Kai Cheng, Hao Chen, Wei Yin, Guangkai Xu, and Xuejin Chen. Exploiting correspondences with all-pairs correlations for multi-view depth estimation. *arXiv preprint arXiv:2205.02481*, 2022.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Weilun Feng, Haotong Qin, Mingqiang Wu, Chuanguang Yang, Yuqi Li, Xiangqi Li, Zhulin An, Libo Huang, Yulun Zhang, Michele Magno, et al. Quantized visual geometry grounded transformer. *arXiv preprint arXiv:2509.21302*, 2025.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 753–762, 2025.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3203–3211, 2025.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 406–413, 2014.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–9502, 2024.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3530–3539, 2022.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1925–1934, 2017.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22820–22830, 2025.

Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16695–16705, 2025.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

Tuan Duc Ngo, Jiahui Huang, Seoung Wug Oh, Kevin Blackburn-Matzen, Evangelos Kalogerakis, Chuang Gan, and Joon-Young Lee. Dage: Dual-stream architecture for efficient and fine-grained geometry estimation. *arXiv preprint arXiv:2603.03744*, 2026.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20133–20143, 2023.

Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17060–17069, 2025.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, volume 1, pp. 519–528. IEEE, 2006.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3531–3539, 2021.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2930–2937, 2013.

Zhijian Shu, Cheng Lin, Tao Xie, Wei Yin, Ben Li, Zhiyuan Pu, Weize Li, Yao Yao, Xun Cao, Xiaoyang Guo, et al. Litevggt: Boosting vanilla vggt via geometry-aware cached token merging. *arXiv preprint arXiv:2512.04939*, 2025.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Xiaolou Sun, Wufei Si, Wenhui Ni, Yuntian Li, Dongming Wu, Fei Xie, Runwei Guan, He-Yang Xu, Henghui Ding, Yuan Wu, et al. Autofly: Vision-language-action model for uav autonomous navigation in the wild. *arXiv preprint arXiv:2602.09657*, 2026.

Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pp. 298–372. Springer, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *IEEE Robotics and Automation Letters*, 5(2):3307–3314, 2020.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Pi3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025d.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22378–22389, 2024.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021.

Guangkai Xu, Hua Geng, Huanyi Zheng, Songyi Yin, Yanlong Sun, Hao Chen, and Chunhua Shen. Carve: Dissecting core components for accurate and resolution-enhanced visual geometry estimation, 2025. URL https://openreview.net/forum?id=WFpZjEogVH.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1790–1799, 2020.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Hao Yu, Haotong Lin, Jiawei Wang, Jiaxin Li, Yida Wang, Xueyang Zhang, Yue Wang, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. Infinidepth: Arbitrary-resolution and fine-grained depth estimation with neural implicit fields. *arXiv preprint arXiv:2601.03252*, 2026.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. Infinitevggt: Visual geometry grounded transformer for endless streams. *arXiv preprint arXiv:2601.02281*, 2026.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21936–21947, 2025.

Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European Conference on Computer Vision*, pp. 523–542. Springer, 2022.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

A Training Details

We train Hybrid-VGGT end-to-end on a distributed cluster comprising 64 NVIDIA A800 GPUs. The model processes temporal sequences at a native high resolution of 1036p, with each GPU handling a randomly sampled batch of 2 to 24 frames per scene. The entire framework is trained for 160 epochs with 1,000 iterations per epoch, resulting in a total of 160,000 training steps. Optimization is performed using AdamW Loshchilov & Hutter (2017) with a weight decay of 0.05. We employ a base learning rate of 2×10^{-5}, which is specifically scaled down to 4×10^{-6} for the patch embeddings. The learning rate is modulated by a cosine annealing schedule, following a 4,000-iteration linear warmup. During training, we apply standard data augmentations including color jittering, random spatial scaling, and center cropping to improve generalization. Finally, to ensure training stability and memory efficiency during high-resolution processing, we utilize Automatic Mixed Precision (AMP), enable TF32 execution, and apply gradient clipping with a maximum L_2 norm of 1.0 across all modules.

To guarantee wide applicability and robust zero-shot generalization, our training corpus aggregates 16 diverse publicly available datasets. This massive mixture encompasses both indoor and outdoor environments, ranging from noise-free synthetic renderings to complex real-world captures. Specifically, the training data includes BlendedMVS (Train Set) Yao et al. (2020), Hypersim Roberts et al. (2021), GTA-SfM Wang & Shen (2020), Replica Straub et al. (2019), TartanAir Wang et al. (2020), Virtual KITTI 2 Cabon et al. (2020), ASE Avetisyan et al. (2024), ADT Pan et al. (2023), MegaDepth Li & Snavely (2018), ScanNet++ Yeshwanth et al. (2023), Taskonomy Zamir et al. (2018), WildRGB-D Xia et al. (2024), MapFree Arnold et al. (2022), CO3Dv2 (Train Set) Reizenstein et al. (2021), Mapillary Neuhold et al. (2017), and DL3DV Ling et al. (2024).

To mitigate label noise and prevent optimization bias across such a heterogeneous collection, we apply rigorous dataset-specific geometric filtering. This includes applying dynamic object masking to exclude non-rigid motion and utilizing depth quantile clipping to remove sensor outliers and boundary artifacts. Furthermore, we employ a probability-weighted sampling strategy to balance the exposure of diverse scene structures, scale variations, and camera trajectories during training, ensuring that the model does not overfit to the most heavily represented datasets.

B Evaluation Details

In this section, we detail our rigorous zero-shot evaluation protocol, defining the metrics, datasets, and specific resolution configurations utilized to ensure a fair comparison between Hybrid-VGGT and existing baselines. While conventional baselines typically operate at a standardized input resolution of 518p, our framework natively infers at 1036p. To guarantee fairness while accurately validating our architectural capabilities, we adopt a task-specific evaluation protocol based on the intrinsic mathematical properties of each downstream metric.

B.1 Camera Pose Recovery

Camera pose represents a global spatial transformation, making its evaluation fundamentally decoupled from the internal 2D image resolution. Therefore, baseline methods calculate poses directly from their native 518p inferences, while we evaluate from our 1036p outputs. This comparison is strictly fair because the pairwise pose error ϵ_{pose} is calculated in the continuous $SE(3)$ space:

$$\epsilon_{pose} = \max(\Delta R, \Delta t) \tag{7}$$

This metric depends exclusively on the rigid rotation (ΔR) and translation (Δt) distance errors after Procrustes alignment. Because it computes a single global transformation matrix for the camera trajectory rather than evaluating a per-pixel grid, the metric is mathematically agnostic to the inference

resolution. We report the Area Under the Curve at a 30° threshold (AUC_{30} $\uparrow$) for Co3Dv2 Reizenstein et al. (2021) and RealEstate10K Zhou et al. (2018), and Absolute Trajectory Error (ATE $\downarrow$) alongside Relative Pose Errors for both translation (RPE-t $\downarrow$) and rotation (RPE-r $\downarrow$) for the TUM-dynamics Zhao et al. (2022) and ScanNet++ Yeshwanth et al. (2023) benchmarks.

B.2 Multi-View Depth Estimation

Unlike global pose, dense depth estimation inherently requires strict pixel-to-pixel correspondence with the ground truth. To rigorously assess the preservation of fine-grained structural boundaries, all depth evaluations are standardized against the 1036p high-resolution Ground Truth (GT). Consequently, baseline methods must bilinearly upsample their 518p depth predictions to 1036p prior to evaluation. This upsampling process inevitably introduces interpolation artifacts and blurs geometric edges. In contrast, Hybrid-VGGT directly outputs high-fidelity 1036p depth maps, ensuring geometric integrity without relying on post-processing upsampling. Following CUT3R Wang et al. (2025b), we quantify performance using the Absolute Relative Error (Abs Rel $\downarrow$):

$$\text{Abs Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i - \hat{d}_i|}{\hat{d}_i} \tag{8}$$

where d_i and $\hat{d}_i$ are the ground-truth and predicted depths at pixel i, respectively. We also report the threshold accuracy (δ_1 $\uparrow$), denoting the percentage of pixels satisfying $\max(d_i/\hat{d}_i, \hat{d}_i/d_i) < 1.25$, evaluated across the Sintel Bozic et al. (2021), ETH3D Schops et al. (2017), 7-Scenes Shotton et al. (2013), BlendedMVS Yao et al. (2020), and ScanNet++ Yeshwanth et al. (2023) benchmarks.

B.3 3D Point Map Reconstruction

Unprojected 3D point maps are evaluated based on their absolute geometric structure, making them independent of the 2D resolution from which they were derived. Both the 518p baselines and our 1036p framework are evaluated directly on their native outputs. Following standard protocols, predictions are globally aligned to the ground truth via Umeyama Sim(3) alignment and Iterative Closest Point (ICP) refinement. The structural fidelity is then rigorously evaluated using Accuracy (ACC $\downarrow$), Completeness (Comp $\downarrow$), and Normal Consistency (NC $\uparrow$) across standard datasets including DTU Jensen et al. (2014), 7-Scenes Shotton et al. (2013), and ETH3D Schops et al. (2017).

Additionally, specifically for our internal ablation studies, we quantify spatial deviation using the 3D Point Map Error (XYZ):

$$E_{XYZ} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{P}_i - \hat{\mathbf{P}}_i\|_2 \tag{9}$$

where $\mathbf{P}_i, \hat{\mathbf{P}}_i \in \mathbb{R}^3$ denote the continuous 3D spatial coordinates. Because these ablation experiments exclusively compare architectural variants of our own framework, all E_{XYZ} evaluations are consistently conducted at the native 1036p resolution, naturally eliminating any resolution bias.

C More Results

Furthermore, Fig. 4 presents additional depth map visualizations to complement the 3D point cloud results. The zoomed-in patches clearly illustrate the limitations of existing baselines: VGGT frequently yields over-smoothed depth boundaries, while Pi3 introduces noticeable blocky artifacts near object edges. In contrast, Hybrid-VGGT consistently reconstructs sharp and precise depth discontinuities. As shown across various indoor scenes, our model effectively recovers thin structures and fine-grained geometric contours, such as delicate chair legs and thin lamp poles. These qualitative improvements explicitly validate the efficacy of our asymmetric dual-branch architecture in preserving high-frequency details without compromising the global scene context.

Fig. 5 provides additional qualitative comparisons of dense point cloud reconstructions across diverse environments. As highlighted in the zoomed-in regions, standard baselines struggle with high-frequency spatial features: VGGT tends to produce over-smoothed surfaces, while Pi3 exhibits noticeable structural artifacts when scaling to complex geometries. In contrast, our Hybrid-VGGT

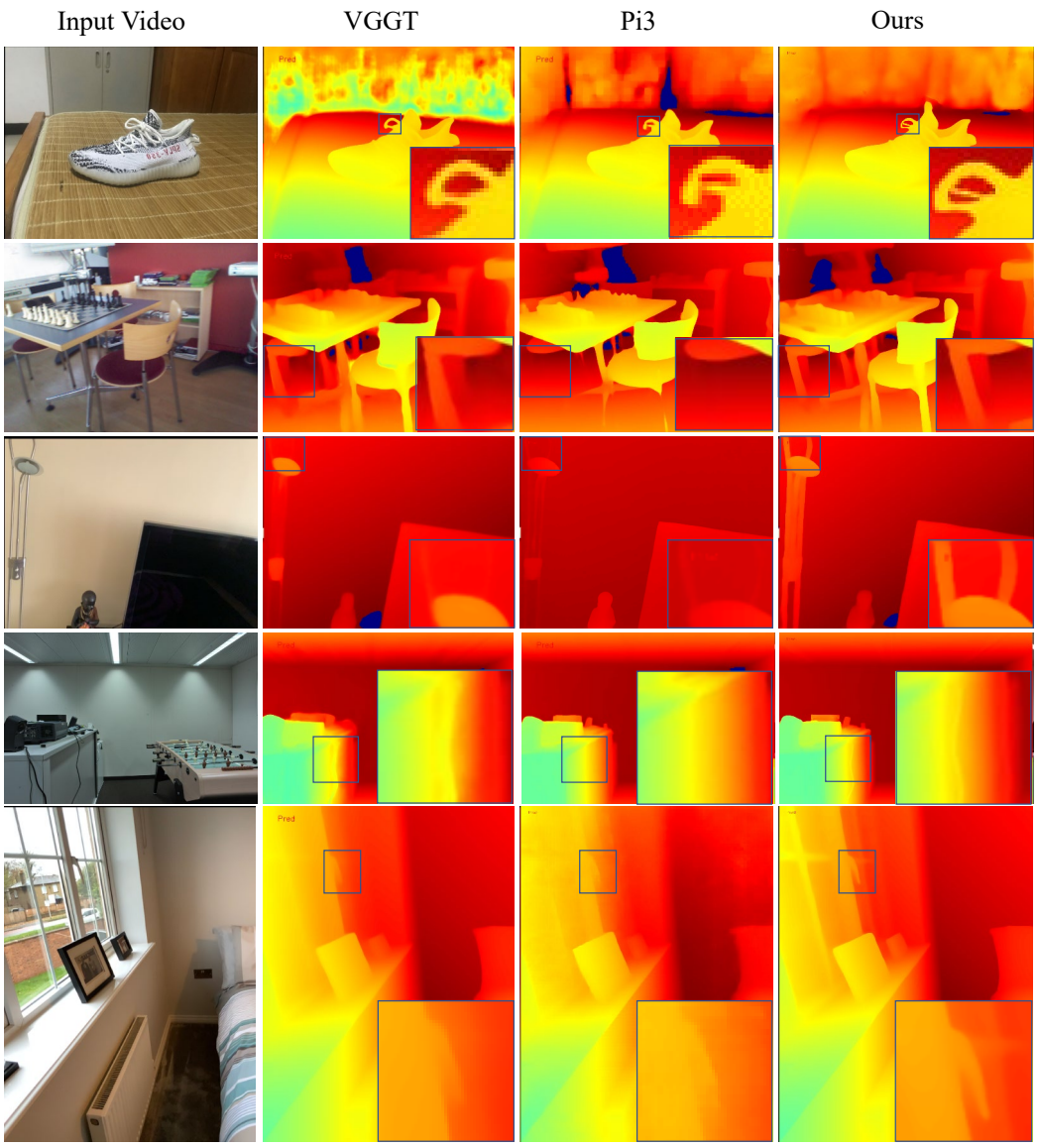


Figure 4: Additional Qualitative Depth Map Comparisons. We compare the dense depth predictions of our Hybrid-VGGT against VGGT and Pi3. The insets highlight regions with complex geometric boundaries and thin structures, demonstrating our method's superiority in recovering sharp depth discontinuities.

robustly preserves fine-grained geometric details. The visual results demonstrate our model's effectiveness in recovering thin structures (e.g., window blinds, church spires) and high-frequency textures. These improvements explicitly validate the capacity of our Hybrid Detail Extractor (HDE) to maintain structural precision and sharp depth boundaries at high resolutions.

D Limitations and Future Work

While Hybrid-VGGT establishes state-of-the-art results for high-resolution geometry estimation, several limitations remain. Primarily, our framework implicitly assumes static scene layouts, making the reconstruction of highly dynamic environments with non-rigid deformations a formidable challenge. Furthermore, extreme physical scenarios involving textureless regions and transparent



Figure 5: Additional Qualitative Point Cloud Comparisons. We compare the dense 3D reconstruction results of our Hybrid-VGGT against VGGT and Pi3 across diverse environments. The insets highlight specific high-frequency regions.

glass can introduce localized depth ambiguities. Future work will explore integrating explicit 4D motion priors and material reasoning to elevate reconstruction fidelity in these complex edge cases.